

CLAIMS

What is claimed is:

1. A computer-implemented method comprising:
applying, by a computer, a plurality of task-specific machine learning models on an inbound audio signal having one or more speaker-independent characteristics to extract a plurality of speaker-independent embeddings for the inbound audio signal;
extracting, by the computer, a deep phoneprint (DP) vector for the inbound audio signal based upon the plurality of speaker-independent embeddings extracted for the inbound audio signal; and
applying, by the computer, one or more post-modeling operations on the plurality of speaker-independent embeddings extracted for the inbound audio signal to generate one or more post-modeling outputs for the inbound audio signal.
2. The method according to claim 1, further comprising
extracting, by the computer, a plurality of features from a training audio signal, the plurality of features including at least one of: a spectro-temporal feature and metadata associated with the training audio signal; and
extracting, by the computer, the plurality of features from the inbound audio signal.
3. The method according to claim 2, wherein the metadata includes at least one of: a microphone-type used to capture the training audio signal, a device type from which the training audio signal originated, a codec type applied for compression and decompression of the training audio signal for transmission, a carrier from which the training audio signal was originated, a spoofing service used to change a source identifier associated with the training audio signal, a geography associated with the training audio signal, a network type from which the training audio signal is originated, and an audio event indicator associated with the training audio signal; and
wherein one or more classifications of the one or more post-modeling outputs include at least one of: the microphone-type, the device type, the codec type, the carrier, the spoofing service, the geography, the network type and the audio event indicator.
4. The method according to claim 1, wherein a post-modeling operation of the one or more post-modeling operations includes at least one of a classification operation and a regression operation.

5. The method according to claim 1, wherein a task-specific machine learning model comprises at least one of a neural network and a Gaussian mixture model.

6. The method according to claim 1, further comprising training, by the computer, the plurality of task-specific machine learning models by applying each of the plurality of task-specific machine learning models on a plurality of training audio signals having the one or more speaker-independent characteristics.

7. The method according to claim 6, wherein training a task-specific machine learning model includes:

extracting, by the computer, one or more first spectro-temporal features for speech portions of a training audio signal and one or more second spectro-temporal features for non-speech portions of the training audio signal;

applying, by the computer, one or more modeling layers of the task-specific machine learning model on the one or more first spectro-temporal features and the one or more second spectro-temporal features to generate a first speaker-independent embedding and a second speaker-independent embedding;

generating, by the computer, predicted output data for the training audio signal by applying one or more post-modeling layers of the task-specific machine learning model on the first speaker-independent embedding and on the second speaker-independent embedding; and

adjusting, by the computer, one or more hyper-parameters of the one or more modeling layers of the task-specific machine learning model by executing a loss function using the predicted output data and expected output data indicated by one or more labels associated with the training audio signal.

8. The method according to claim 6, wherein training a task-specific machine learning model includes:

extracting, by the computer, one or more first spectro-temporal features for speech portions of a training audio signal and one or more second spectro-temporal features for non-speech portions of the training audio signal;

applying, by the computer, one or more shared modeling layers of one or more task-specific machine learning models on the one or more first spectro-temporal features and the one or more second spectro-temporal features to generate a first speaker-independent embedding and a second speaker-independent embedding;

generating, by the computer, predicted output data for the training audio signal by applying one or more post-modeling layers of the task-specific machine learning model on the first speaker-independent embedding and on the second speaker-independent embedding; and
adjusting, by the computer, one or more hyper-parameters of the task-specific machine learning model by executing a shared loss function of the one or more task-specific machine learning models using the predicted output data and expected output data indicated by one or more labels associated with the training audio signal.

9. The method according to claim 6, wherein training the plurality of task-specific machine learning models includes:

extracting, by the computer, one or more spectro-temporal features for a substantial portion of a training audio signal;

applying, by the computer, a task-specific machine learning model on the one or more spectro-temporal features of the training audio signal to generate a speaker-independent embedding;

generating, by the computer, predicted output data for the training audio signal by applying one or more post-modeling layers of the task-specific machine learning model on the speaker-independent embedding for the substantial portion of the training audio signal; and

adjusting, by the computer, one or more hyper-parameters of the task-specific machine learning model by executing a loss function using the predicted output data and expected output data indicated by one or more labels associated with the training audio signal.

10. The method according to claim 1, wherein a task-specific machine learning model comprises at least one of a convolutional neural network, recurrent neural network, and a fully connected neural network.

11. The method according to claim 1, further comprising:

executing, by the computer, a voice activity detection (VAD) operation on a training audio signal, thereby generating one or more speech portions for the training audio signal and one or more non-speech portions for the training audio signal; and

executing, by the computer, the VAD operation on the inbound audio signal, thereby generating the one or more speech portions for the inbound audio signal and the one or more non-speech portions for the inbound audio signal.

- 12.** The method according to claim 1, further comprising:
- extracting, by the computer, a plurality of enrolled speaker-independent embeddings and DP vectors for an enrolled audio source by applying the plurality of task-specific machine learning models on a plurality of enrollment audio signals associated with the enrolled audio source; and
 - extracting, by the computer, the DP vector for the enrolled audio source based upon the plurality of enrolled speaker-independent embeddings,
 - wherein the computer generates the one or more post-modeling outputs using the DP vector for the enrolled audio source and the DP vector for the inbound audio signal.
- 13.** The method according to claim 12, further comprising determining, by the computer, an authentication classification for the inbound audio signal based upon one or more similarity scores between the DP vector for the enrolled audio source and the DP vector for the inbound audio signal.
- 14.** A system comprising:
- a database comprising non-transitory memory configured to store a plurality of audio signals having one or more speaker-independent characteristics; and
 - a server comprising a processor configured to:
 - apply a plurality of task-specific machine learning models on an inbound audio signal having the one or more speaker-independent characteristics to extract a plurality of speaker-independent embeddings for the inbound audio signal;
 - extract a deep phoneprint (DP) vector for the inbound audio signal based upon the plurality of speaker-independent embeddings extracted for the inbound audio signal; and
 - apply one or more post-modeling operations on the plurality of speaker-independent embeddings extracted for the inbound audio signal to generate one or more post-modeling outputs for the inbound audio signal.
- 15.** The system according to claim 14, wherein the server is further configured to:
- extract a plurality of features from a training audio signal, the plurality of features including at least one of: a spectro-temporal feature and metadata associated with the training audio signal; and
 - extract the plurality of features from the inbound audio signal.

16. The system according to claim 15, wherein the metadata includes at least one of: a microphone-type used to capture the training audio signal, a device type from which the training audio signal originated, a codec type applied for compression and decompression of the training audio signal for transmission, a carrier through from which the training audio signal was originated, a spoofing service used to change a source identifier associated with the training audio signal, a geography associated with the training audio signal, a network type from which the training audio signal is originated, and an audio event indicator associated with the training audio signal; and

wherein one or more classifications of the one or more post-modeling outputs include at least one of: the microphone-type, the device type, the codec type, the carrier, the spoofing service, the geography, the network type, and the audio event indicator.

17. The system according to claim 14, wherein a post-modeling operation of the one or more post-modeling operations includes at least one of a classification operation and a regression operation.

18. The system according to claim 14, wherein a task-specific machine learning model comprises at least one of a neural network and a Gaussian mixture model.

19. The system according to claim 14, wherein the server is further configured to train the plurality of task-specific machine learning models by applying each of the plurality of task-specific machine learning models on a plurality of training audio signals having the one or more speaker-independent characteristics.

20. The system according to claim 19, wherein when training a task-specific machine learning model the server is further configured to:

extract one or more first spectro-temporal features for speech portions of a training audio signal and one or more second spectro-temporal features for non-speech portions of the training audio signal;

apply one or more modeling layers of the task-specific machine learning model on the one or more first spectro-temporal features and the one or more second spectro-temporal features to generate a first speaker-independent embedding and a second speaker-independent embedding;

generate predicted output data for the training audio signal by applying one or more post-modeling layers of the task-specific machine learning model on the first speaker-independent embedding and on the second speaker-independent embedding; and

adjust one or more hyper-parameters of the one or more modeling layers of the task-specific machine learning model by executing a loss function using the predicted output data and expected output data indicated by one or more labels associated with the training audio signal.

21. The system according to claim 19, wherein when training a task-specific machine learning model the server is further configured to:

extract one or more first spectro-temporal features for speech portions of a training audio signal and one or more second spectro-temporal features for non-speech portions of the training audio signal;

apply one or more shared modeling layers of one or more task-specific machine learning models on a plurality of features on the one or more first spectro-temporal features and the one or more second spectro-temporal features to generate a first speaker-independent embedding and a second speaker-independent embedding;

generate predicted output data for the training audio signal by applying one or more post-modeling layers of the task-specific machine learning model on the first speaker-independent embedding and on the second speaker-independent embedding; and

adjust one or more hyper-parameters of the task-specific machine learning model by executing a shared loss function of the one or more task-specific machine learning models using the predicted output data and expected output data indicated by one or more labels associated with the training audio signal.

22. The system according to claim 19, wherein when training the plurality of task-specific machine learning models the server is further configured to:

one or more spectro-temporal features for a substantial portion of a training audio signal;

apply a task-specific machine learning model on the one or more spectro-temporal features of the training audio signal to generate a speaker-independent embedding;

generate predicted output data for the training audio signal by applying one or more post-modeling layers of the task-specific machine learning model on the speaker-independent embedding for the substantial portion of the training audio signal; and

adjust one or more hyper-parameters of the task-specific machine learning model by executing a loss function using the predicted output data and expected output data indicated by one or more labels associated with the training audio signal.

23. The system according to claim 14, wherein a task-specific machine learning model comprises at least one of a convolutional neural network, a recurrent neural network, and a fully connected neural network.

24. The system according to claim 14, wherein the server is further configured to:
execute a voice activity detection (VAD) operation on a training audio signal, thereby generating one or more speech portions for the training audio signal and one or more non-speech portions for the training audio signal; and
execute the VAD operation on the inbound audio signal, thereby generating the one or more speech portions for the inbound audio signal and the one or more non-speech portions for the inbound audio signal.

25. The system according to claim 14, wherein the server is further configured to:
extract a plurality of enrolled speaker-independent embeddings and DP vectors for an enrolled audio source by applying the plurality of task-specific machine learning models on a plurality of enrollment audio signals associated with the enrolled audio source; and
extract the DP vector for the enrolled audio source based upon the plurality of enrolled speaker-independent embeddings,
wherein the server generates the one or more post-modeling outputs using the DP vector for the enrolled audio source and the DP vector for the inbound audio signal.

26. The system according to claim 25, wherein the server is further configured to:
determine an authentication classification for the inbound audio signal based upon one or more similarity scores between the DP vector for the enrolled audio source and the DP vector for the inbound audio signal.